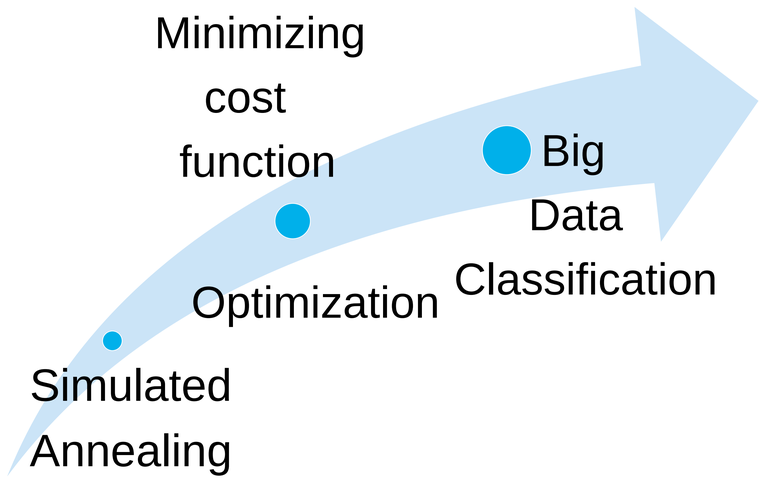
Global Optimization for Big Data Classification Using Simulated Annealing

Angli Zhao College of Liberal Arts University of Minnesota Twin Cities Mentor: Charles Geyer, Xiaotong Shen

Overview

Optimization remains a hot topic in science and engineering due to its wide application, which in turns brings in lots of algorithms to solve such problem. **Simulated annealing** is a generic method of finding the global optimum of functions with many local minima. Unlike deterministic local search algorithms that always go downhill and thus get trapped in any local minimum, simulated annealing and other adaptive random search algorithms make random "proposals", which sometimes accepts uphill steps with probability that decreases with a parameter called "temperature" so that it will not get trapped and keep searching. Simulated annealing methods in the literature do not adjust the proposal distribution as the temperature changes, resulting in almost all proposals being rejected as the temperature goes to zero. Here we show that simulated annealing has better performance if the **proposal variance** is a **linear function** of **temperature** because this keeps the proposal acceptance rate about the same.

Classification and prediction using **big data** is another hot topic. All widely used methods use optimization and need to find global optima. Many methods minimize convex functions and hence automatically find global minima, however, such methods do not have as good classification or prediction accuracy as those minimizing non-convex functions. There are many **non-convex** methods in the literature, but we are most interested in one called **psi-learning**. And we apply simulated annealing to find global minima for it.



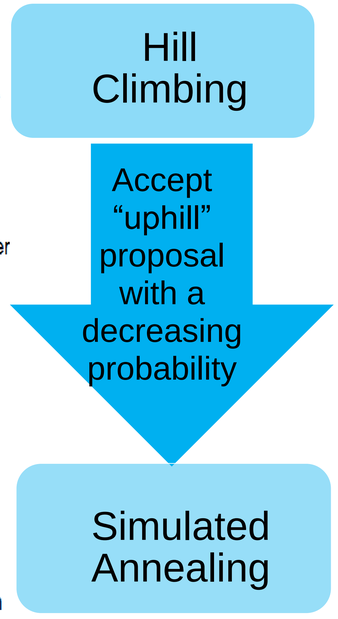
Simulated Annealing

The most intuitive method of finding minimum is hill climbing.

Given a function f(x) and a random staring point x_0, we can find its minimum by move Δx at certain times. For example, let $\Delta x > 0$, if $f(x_0+\Delta x) < f(x_0)$, we accept this move and $(x_0+\Delta x)$ will be our new staring point, otherwise, we reject this move, eventually we can reach x_k where $f(x_k)$ is smaller than the value of any point in its neighborhood. However, we may not conclude that $f(x_k)$ is the minimum of the function f(x) because there may exist x_n, which is outside the neighborhood of x_k, yet $f(x_n)$ is smaller than $f(x_k)$. In this case, $f(x_k)$ is the local minimum in its neighborhood while $f(x_n)$ is the global minimum of function f(x).

Generally speaking, we can hardly find the global minimum by hill climbing. That's why we use simulated annealing.

If the move which traps us with local minimum is allowed with certain probability, there are possibilities for us jumping out of the local minimum and reaching the global minimum in the end.

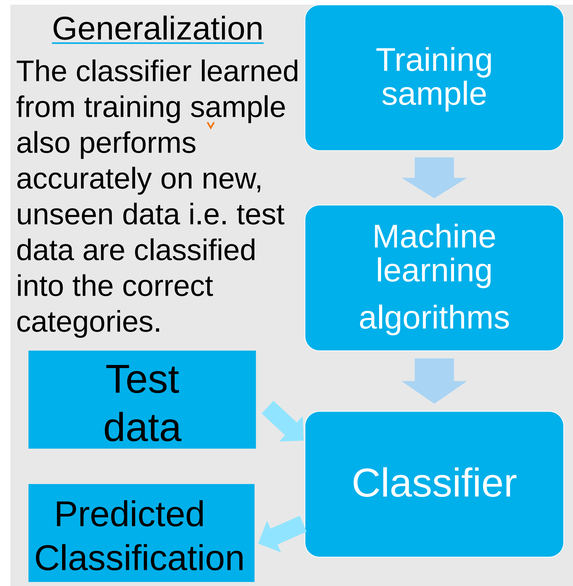


The index to control is the probability ("cooling schedule") and Δx ("proposal")

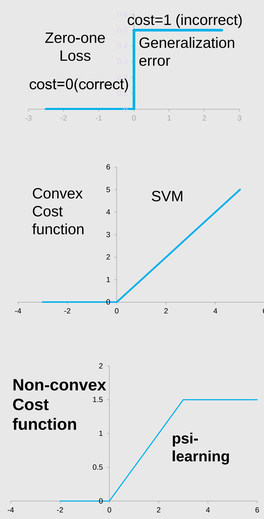
Classification

Generalization

The classifier learned from training sample also performs accurately on new, unseen data i.e. test data are classified into the correct categories.



Cost function



Conclusion

For simulated annealing, initially, we need a larger probability (higher temperature) so that it is more likely to accept the unqualified proposal and help us get rid of the local minimum. Then the probability of accepting $f(x_0+\Delta x)$, when it is not smaller than $f(x_0)$, should decrease gradually (cool down the temperature) to improve the solution. What we need to figure out is how much the temperature and x should vary at each time to guarantee a global optimization, i.e. how to find the appropriate "cooling schedule" and "proposal".

It turns out that if the proposal distribution is normal with mean zero, then proposal variance should be a linear function of the temperature to keep a constant acceptance rate of worse solutions. This is a new idea in since traditional simulated annealing does not vary the proposal distribution with temperature. We have done the proof by referring to the literature of Metropolis methods, which can be treated as a simulated annealing procedure without cooling schedule. We also show that it improves the performance of finding the global minimum.